ON ANALYSIS OF VARIANCE COMPUTING PACKAGE OUTPUT FOR UNBALANCED DATA FROM FIXED EFFECTS MODELS WITH NESTED FACTORS.

Shayle R. Searle

Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-1203-M April 1993

ABSTRACT

Explanations are offered for some of the idiosyncrasies evident in computer output of sums of squares of unbalanced data described by Dallal (1992).

1. INTRODUCTION

Dallal (1992) presents an interesting example (brought to my attention by Dr. John Randall of the University of Stellenbosch) of some difficult-to-understand sums of squares produced by computing packages. The data are those of Table 1, a 2-way crossed classification of two rows (factor A) and three columns (factor B) with the A-by-B cells in row 1 having three observations and those of row 2 having six observations. Dallal reports on analyzing these data with two different computing packages, SAS GLM using its Type III sums of squares and SPSS MANOVA using its Unique sums of squares.

Table 1. Dallal's Data

	B1	B2	B3
A1	3.81	3.42	3.55
	4.64	3.57	3.71
	4.09	3.55	3.66
A2	0.22	0.36	0.37
	0.33	0.27	0.31
	0.36	0.26	0.28
	1.08	0.83	0.70
	1.33	0.90	0.89
	1.15	0.93	0.93

With each package the analysis was done with two different (but seemingly statistically similar) models. With each package the same sums of squares were not obtained for the two models despite their apparent similarity. This note discusses several aspects of the disparity when using SAS GLM.

2. THE MODELS

What shall be called the standard (overparameterized) model, which it is, is that of main effects A and B and interaction A * B. A suitable model equation for this is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} , \tag{1}$$

where y_{ijk} is the k'th observation in the cell defined by the i'th level of A and j'th level of B. μ is a general mean, α_i is the A-effect, β_j is the B-effect, $(\alpha\beta)_{ij}$ is the interaction effect and e_{ijk} is the residual error.

An alternative model equation is that for the cell means model

$$y_{ijk} = \mu_{ij} + e_{ijk} , \tag{2}$$

where μ_{ij} is the population cell mean for cell (i, j) and y_{ij} is the random variable for a datum in cell (i, j). For both (1) and (2) applied to Table 1,

$$i = 1, 2 \text{ and } j = 1, 2, 3, \text{ with } k = 1, 2, 3 \text{ for } i = 1, \text{ and } k = 1, 2, \cdots, 6 \text{ for } i = 2 . \tag{3}$$

In the case of model equation (1) there is no contention about calculating sums of squares such as those for A adjusted for μ, $R(A \mid \mu)$, or for B adjusted for A and μ, $R(B \mid \mu, A)$ and for interaction adjusted for main effects, $R(A * B \mid \mu, A, B)$. These are often referred to as sequential sums of squares when the factors are treated by a computing package in the sequence A, B and A * B. And a similar sequence can be B, A and A * B. There is no argument about such sums of squares; and Dallal implicitly concurs. Moreover, they are the SAS Type I sums of squares.

The second model that he uses, which we shall call the C-model, is where he labels the observation within each A-by-B cell as a factor C within A * B; thus it has 3 levels in each cell of A1 in Table 1, and 6 levels in each cell of A2 of Table 1. And every level of C within A * B has but one observation. For this, a suitable model equation is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{ij:k} . \tag{4}$$

The indices here are the same as in (3) for (1) and (2). Notice that in (4) there is no error term because there is no opportunity for one because in each level of C, which is nested within A*B, there is only one observation.

Similar to the cell means model equation (2) being an alternative to the overparameterized model equation (1), we will also consider

$$y_{ijk} = \mu_{ij} + \gamma_{ij:k} \tag{5}$$

as an alternative to (4). Effectively, (5) is just (2) with the $\gamma_{ij:k}$ term of (4) in place of e_{ijk}.

The difficult-to-understand sums of squares arose as between the two models (1) and (4) for sums of squares other than those of the form described in the paragraph following (3). In particular, it was the SAS Type III sum of squares for B that was not the same for the C-model (4) as for the standard model (1). Yet one would expect them to be the same because, apart from a simple labelling change, the two models are the same. In (4) the only change from (1) is labelling each ϵ_{ijk} of (1) as $\gamma_{ij:k}$ of (4). And this should cause no difference in analysis-of-variance style calculations.

3. SAS TYPE III SUMS OF SQUARES

Cross-classified data: no nested factors

It is well known for cross-classified fixed effects models with all-cells-filled data that the SAS Type III sums of squares are those of Yates (1934) weighted squares of means analysis. They can also be described in two other ways. First, they are the sums of squares, for all-cells-filled data, for overparameterized Σ-restricted models. Such a model is (1) with the following restrictions imposed on its parameters.

$$\Sigma_i \alpha_i = 0, \quad \Sigma_j \beta_j = 0, \quad \Sigma_i (\alpha\beta)_{ij} = 0 \ \forall \ \mathrm{j} \quad \text{and} \quad \Sigma_j (\alpha\beta)_{ij} = 0 \ \forall \ \mathrm{i} \ . \tag{6}$$

Second, the Type III sums of squares can be described in terms of the hypotheses they test. For model equation (1) this is as follows:

$$\begin{aligned} &\text{Type III} \quad \text{SS(A)} \ \text{tests} \ \mathrm{H}: \alpha_i + \Sigma_j (\alpha\beta)_{ij} \big/ 3 \ \ \textit{equal} \ \forall \ \mathrm{i} \ , \\ &\text{Type III} \quad \text{SS(B)} \ \text{tests} \ \mathrm{H}: \beta_j + \Sigma_i (\alpha\beta)_{ij} \big/ 2 \ \ \textit{equal} \ \forall \ \mathrm{j} \ . \end{aligned} \tag{7}$$

By "a sum of squares tests H" we mean that in a fixed effects model, when that sum of squares is converted to its mean square and divided by the estimated residual variance, then that ratio is an F-

statistic suitable for testing H. Rather than use this long description of F it is clearly easier to use "a sum of squares tests H" when discussing different sums of squares and their utility.

The hypotheses in (7) are hypotheses of equality of main effects in the presence of averaged interactions; and if one uses the Σ-restrictions of (6) those average interactions disappear from (7). However, for purposes of presenting the hypotheses of different sums of squares, cell means models, such as (2), are an informative vehicle. Thus for (2) the hypotheses of (7) are

$$\begin{aligned} &\text{Type III} \quad \text{SS(A)} \ \text{tests} \ \mathrm{H}: \bar{\mu}_{i\cdot} \ \textit{all equal}, \\ &\text{Type III} \quad \text{SS(B)} \ \text{tests} \ \mathrm{H}: \bar{\mu}_{\cdot j} \ \textit{all equal}. \end{aligned} \tag{8}$$

These are hypotheses of equality of means of cell means: $\bar{\mu}_{i\cdot}$ for Table 1 is $\mu_{i\cdot} = (\mu_{i1} + \mu_{i2} + \mu_{i3})/3$ and $\bar{\mu}_{\cdot j} = (\mu_{1j} + \mu_{2j})/2$.

The presentation in (8) of hypotheses in terms of means of cell means extends, for all-cells-filled data, quite naturally beyond the 2-way classification of Table 1. For example, for a 3-way crossed classification, as discussed in Searle (1987, Section 10.2),

$$\begin{aligned} &\text{Type III} \quad \text{SS(A)} \quad \text{tests} \ \mathrm{H}: \bar{\mu}_{i\cdot\cdot} \ \textit{all equal} \\ &\text{and} \\ &\text{Type III} \quad \text{SS(AB)} \ \text{tests} \ \mathrm{H}: \bar{\mu}_{ij\cdot} - \bar{\mu}_{i'j\cdot} - \bar{\mu}_{ij'\cdot} + \bar{\mu}_{i'j'\cdot} = 0 \quad \forall \ \mathrm{i} \neq \mathrm{i}' \ \text{and} \ \mathrm{j} \neq \mathrm{j}'. \end{aligned} \tag{9}$$

Dallal's results

As shall now be discussed, Dallal's data indicates that when unbalanced all-cells-filled data are from a fixed effects model with a nested factor, then (8) is not necessarily true and neither, therefore, would (9) be true – nor any of its otherwise natural extensions.

The Type III sums of squares reported by Dallal for both the standard model of (1) and the C-model of (4) are shown in Table 2.

Table 2. Type III Sums of Squares for Data of Table 1

	Standard Model, (1)	C-Model (4)
SS(A)	59.11574074	59.11574074
SS(B)	0.78843333	0.75766365
SS(A*B)	0.28218148	0.28218148
SS[C(A*B)]	—	2.57130000
Residual	2.57130000	—

As noted earlier, with the C-model the residual sum of squares is zero and SS[C(A * B)] is the same as the residual in the standard model. The puzzle, though, is why is SS(B) not the same in the C-model as in the standard model? After all, the C-model is no more than the standard model with a factor label, C, attached to the within A-by-B cell observations. Why should that affect SS(B)? Three ideas were explored for trying to explain this difference.

Factor sequencing

A first question was "Why is factor B affected but not A?". There seems no obvious reason for one factor being affected differently from the other. So maybe the sequencing of the main effects was a root cause. But entering the factors as B then A rather than A then B made no difference – as was to be expected on referring to SAS manuals: sequencing of main effects does not affect Type III sums of squares.

Hypotheses tested

A second idea was to look at the SAS GLM output of estimable functions for the Type III sums of squares. This output provides information suitable for constructing hypotheses tested by the sums of squares (e.g., Searle, 1987, Section 12.3). From this it was hoped that one would be able to establish a reason for the hypotheses of the C-model and hence better understand the behavior of Type III sums of squares in the presence of nested factors. Alas, this did not come to pass.

For Table 1, the hypotheses (8) are

SS(A) tests $H: \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot}$, which is $(\mu_{11}+\mu_{12}+\mu_{13})/3-(\mu_{21}+\mu_{22}+\mu_{23})/3 = 0$. (10)

SS(B) tests $H: \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \bar{\mu}_{\cdot 3}$, which is

$$H: \begin{cases} (\mu_{11}+\mu_{21})/2 - (\mu_{12}+\mu_{22})/2 = 0 \\ (\mu_{11}+\mu_{21})/2 - (\mu_{13}+\mu_{23})/2 = 0 \end{cases}. \tag{11}$$

In the 2×3 grid of Table 1 the numbers of observations and the μ_{ij} are as shown in Grids 1 and 2.

Grid 1 n_{ij}

	j = 1	j = 2	j = 3
i = 1	3	3	3
i = 2	6	6	6

Grid 2 μ_{ij}

μ_{11}	μ_{12}	μ_{13}
μ_{21}	μ_{22}	μ_{23}

Then, corresponding to μ_{ij}s of Grid 2, the hypothesis tested by SS(A) for the standard model, as stated in (9), can be represented diagramatically as in Grid 3.

Grid 3. Hypothesis for SS(A) for the Standard Model

$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$

Similarly, the 2-part hypothesis in (10), for SS(B) can be represented by Grids (4a) and (4b).

Grids 4a and 4b. Hypothesis for SS(B) for the Standard Model

Grid 4a

$\frac{1}{2}$	$-\frac{1}{2}$	
$\frac{1}{2}$	$-\frac{1}{2}$	

and

Grid 4b

$\frac{1}{2}$		$-\frac{1}{2}$
$\frac{1}{2}$		$-\frac{1}{2}$

To compare the situation for the C-model with Grids $1-4$, we first recognize that the grid for the C-model is as shown in Grid 5.

Grid 5. n_{ij} for the C-model

	B1			B2			B3		
A1	1	1	1	1	1	1	1	1	1
A2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Then for the C-model, using model equation (5) and taking particular note of the values of the subscript k given in (3),

$$\begin{aligned} \text{SS(A)} \quad \text{tests} \quad \text{H}: &(\mu_{11}+\mu_{12}+\mu_{13})/3+\left(\sum_{k=1}^{3}\gamma_{11:k}+\sum_{k=1}^{3}\gamma_{12:k}+\sum_{k=1}^{3}\gamma_{13:k}\right)\Big/ 9 \\ &-(\mu_{21}+\mu_{22}+\mu_{23})/3+\left(\sum_{k=1}^{6}\gamma_{21:k}+\sum_{k=1}^{6}\gamma_{22:k}+\sum_{k=1}^{6}\gamma_{23:k}\right)\Big/ 18 = 0\,. \end{aligned} \tag{12}$$

To represent this in a grid in the style of Grid 3 [for SS(A) in the standard model], would require a grid like Grid 5. Since that would be rather cumbersome we represent it as Grid 6a for the μ_{ij} of the model equation (5), plus Grid 6b for the $\gamma_{ij:k}$. And for the latter, rather than showing 1/9 three times in the A1 × B1 cell (for example), in accord with (12), we represent the three occurrences of 1/9 as 3 @ 1/9; and so on, for each of the A-by-B cells. Thus Grids 6a and 6b are as shown.

Grids 6a + 6aa. Hypothesis for SS(A) for C-model

Grid 6a: for μ_{ij}

$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$

+

Grid 6aa: for $\gamma_{ij:k}$

$3@\frac{1}{9}$	$3@\frac{1}{9}$	$3@\frac{1}{9}$
$6@\frac{-1}{18}$	$6@\frac{-1}{18}$	$6@\frac{-1}{18}$

It is easily seen that Grid 6a is the same as Grid 3, which is to be expected, since SS(A) is the same for both models, as in Table 2; and Grid 6b is an easily understood add-on.

Now we come to SS(B) for the C-model, which differs from SS(B) for the standard model (see Table 2). As in (10) for the standard model, the hypothesis for SS(B) has two degrees of freedom and so like Grids 4a and 4b, each piece has to be represented in two parts, added together. Thus in Grids 7a and 7aa and 7b and 7bb we have the representation of the hypothesis tested by SS(B) for the C-model.

Grids 7a + 7aa and 7b + 7bb for the hypothesis for SS(B) for the C-model

Grid 7a: for μ_{ij}

$\frac{7}{15}$	$\frac{-7}{15}$	
$\frac{8}{15}$	$\frac{-8}{15}$	

+

Grid 7aa: for $\gamma_{ij:k}$

$3@\frac{7}{45}$	$3@\frac{-7}{45}$	
$6@\frac{4}{45}$	$6@\frac{-4}{45}$	

and

Grid 7b: for μ_{ij}

$\frac{7}{15}$		$\frac{-7}{15}$
$\frac{8}{15}$		$\frac{-8}{15}$

+

Grid 7bb: for $\gamma_{ij:k}$

$3@\frac{7}{45}$		$3@\frac{-7}{45}$
$6@\frac{4}{45}$		$6@\frac{-4}{45}$

The question arising from Grids 7 is "from whence cometh the fractions $\frac{7}{15}$ and $\frac{8}{15}$?". Their counterparts are $\frac{1}{2}$ and $\frac{1}{2}$ in Grids 4 for SS(B) in the standard model. Given the fractions in Grids 7a and 7b, those in 7aa and 7bb are easily understood: they are each one-third of those in 7a and 7b for the A1 level and are one-sixth for the A2 level. But one must wonder where the $\frac{7}{15}$ and $\frac{8}{15}$ come from. They seem to bear no obvious relationship to the three levels of C within each A × B cell in the first row of the grid, and 6 levels in the second row. Failing all else, then, one might contemplate the possibility of looking in SAS documentation, old and new, to see if description of Type III sums of squares might throw light on the subject.

SAS Documentation

Goodnight (1976), in the proceedings of the first SAS users' conference, provides the original description of how the estimable functions corresponding to Type III sums of squares are derived – and it is from these estimable functions that hypotheses can be constructed (e.g., Searle, 1987, Sec. 12.3). Some relevant quotes from the Goodnight article, pertinent to deriving those estimable functions, are as follows.

An effect "e_1 is contained in [another effect] e_2 provided ...

1. Both effects involve the same number of continuous variables and if the number involved is positive, then the names of the continuous variables coincide.
2. e_2 has more class variables than does e_1 and if e_1 has class variables, all class names in e_1 are in e_2." [p.13]

"In order to obtain testable hypotheses given only the general form of estimable functions, several rules based on the nature of the testable hypotheses in balanced designs were developed. The first ... rule is ... that ... the coefficients of all effects not

containing e_1 (except e_1) should be zero. The second rule is that the block of coefficients pertaining to e_1 should all be "free" coefficients or functions of "free" coefficients in that block." [p.18]

"Type III estimable functions for an effect e_1 are computed as follows: First a basis for effect e_1 is formed. In the general form notation, if no "free" coefficients exist outside of the e_1 block then these are the Type III estimable functions for e_1. If "free" coefficients exist outside of the e_1 block then each of these coefficients is equated to a function of the e_1 "free" coefficients in such a way as to make the Type III estimable functions for e_1 orthogonal to all other Type III estimable functions which contain e_1." [p.22]

"Type III estimable functions, have one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant." [p.22]

"When no missing cells exist in a factorial model, Type III SS will coincide with Yates' weighted squares of means technique." [p.22]

"Also, when missing cells exist in a design, the Type III estimable functions for any effect which is contained in another effect, have some rather strange coefficients for the higher order effects." [p.24]

This lifting of isolated sentences out of context is, of course, open to the criticism of being just that; and for any false impressions it conveys, apologies are due and are here offered. Nevertheless, all of the description is in terms of computing. There is only one statistical comment, in the penultimate quote, concerning Yates' weighted squares of means; and that, be it noted, is confined to all-cells-filled data and to factorial designs. Other than that, there is no detailed statistical explanation of what is being computed.

This lack of statistical explanation continues into recent SAS publications. For example, in the 1985 Version 5 Edition of SAS User's Guide: *Statistics* we find (on p.88) the following paragraph.

"**Construction of Type III Hypotheses**

Type III hypotheses are constructed by working directly with the general form of estimable functions. The following steps are used to construct a hypothesis for an effect E1:

1. For every effect in the model except E1 and those effects that contain E1, equate the coefficients in the general form of estimable functions to zero.
   Note: if E1 is not contained in any other effect, this step defines the Type III hypothesis (as well as the Type II and Type IV hypotheses). If E1 is contained in other effects, go on to step 2.
2. If necessary, equate new symbols to compound expressions in the E1 block in order to obtain the simplest form for the E1 coefficients.
3. Equate all symbolic coefficients outside of the E1 block to a linear functions of the symbols in the E1 block in order to make the E1 hypothesis orthogonal to hypotheses associated with effects that contain E1."

And almost exactly the same paragraph, word for word, is found on page 120 of Volume 1 of the 1990 SAS/STAT User's Guide, Version 6, 4th Edition.

Unbalanced data

From all of this we are drawn to the following conclusions concerning Type III sums of squares for nested data, at least on the basis of the data of Table 1 analyzed by the C-model.

Without a good statistical explanation of how the SAS Type III sums of squares are calculated, and with consideration of the hypotheses they test appearing to yield no statistical framework for developing those hypotheses (*vide* the fractions 7/15 and 8/15 in Grids 7), we are driven to consider some other possible reason for the SS(B) value in Table 2 not being the same for the C-model as for the standard model. And (at the suggestion of Dr. Charles McCulloch) we believe it is because of the unbalancedness of the data with respect to the C factor; for the data of Table 1 a very particular aspect of the unbalancedness. In looking at Grid 1 again (and keeping in mind that in the C-model each observation represents one level of C within an A-by-B cell), we see that within each level of A

there is balance with respect to C across the levels of B. Thus in A1 there are three levels of C for each level of B; and in A2 there are six levels of C for each level of B. Apparently, because of this aspect of balancedness, SS(A) in Table 2 is the same for both models. In contrast, this form of balancedness does not apply for SS(B). Within each level of B there is not balance with respect to levels of C over the levels of A. Thus in B1 (and in B2 and B3) there are three levels of C in A1 but six levels of C in A2. And this, apparently, causes SS(B) in Table 2 to be different in the C-model from what it is in the standard model.

Further unbalancedness

To partially test out this argument about balancedness and unbalancedness we dropped one datum from the A1-by-B1 cell so that the n_{ij}-values were as shown in Grid 8.

Grid 8. n_{ij}

2	3	3
6	6	6

For this layout we calculated Table 2. Now SS(A) and SS(B) differed in the C-model from the standard model. And the grids comparable to Grids 6 and Grids 7 were very different. Akin to Grids 6, for SS(A), for example, we got Grids 9.

Grids 9a and 9b: Hypothesis for SS(A) for C-model analysis for Grid 8

Grid 9a: for μ_{ij}

.3191	.3404	.3404
−.3191	−.3404	−.3404

+

Grid 9b: for $\gamma_{ij:k}$

2@.1596	3@.1135	3@.1135
6@−.0532	6@−.0567	6@−.0567

It is clear that the numbers in Grid 9b bear an obvious relationship to those in Grid 9a. But the origin of the latter is far from clear. Reducing them to rational fractions is no help either; for example, .3191 = 1979/6203. And grids like Grid 7 for the hypothesis for SS(B) have even more complicated numbers.

Although the preceding reasoning seems to provide a satisfactory explanation for occurrence of the bothersome difference in the SS(B) values of Table 2, it does not provide statistical explanation of how SS(B) was calculated for the C-model, or of how the hypothesis for that SS(B) was derived. Maybe calculating Table 2 for a smaller data set than Table 1, one with simple, integer data values, could lead to a more informative understanding of the nature of Type III sums of squares in the presence of nested classification. In the meantime, the moral seems to be do not use Type III with nested classifications. This is no great loss insofar as fixed effects models are concerned, because nesting seldom occurs with such models; it occurs mostly in mixed or random models.

4. REFERENCES

Dallal, G.E. (1992). The computer analysis of factorial experiments with nested factors. *The American Statistician*, 46, 240.

Goodnight, J.H. (1976). General linear model procedure. SAS.ONE, Proceedings of the First International Users Conference, Kissimee, Florida, 1-39.

Searle, S.R. (1987). *Linear Models for Unbalanced Data.* John Wiley & Sons, New York.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association* **29**, 51-66.

ADDENDUM

After submitting the preceding paper to *The American Statistician*, its editors asked me to review and comment briefly on replies that they had received from software houses to *The American Statistician*'s request to them for reactions to the Dallal letter. This addendum is my response to that editorial request. The dates following the software names are the dates of their letters.

1. BMDP (2/11/93) Two models are used, defining effects in their terms by weights. In the first model "the six cells defined by the combination of factors A and B are given equal weights". Their second model "creates 27 cells with equal weights" which they write "changes the parameter estimation procedure for the main effects and interaction effects". I do not under-stand this; and less still do I understand their output being the same as for the C-model in Table 1 except their SS(B) is different (.571652) from both the value shown in Table 1 for the standard model and for the C-model.

What BMDP calls Models III and IV are then as described in Table A1. BMDP comments that Model III is equivalent to Model II and Model IV to Model I. These statements of equivalence are correct insofar as the arithmetic results are concerned (see Table A1), but they are not correct concerning which label, Residual or SS[C(A * B)], is attached to the value 2.571.

Frankly, I find these descriptions of analysis of variance calculations given in general terms of weights to be dissatisfyingly non-specific. For example, of themselves, they give no hint of what hypotheses they test. Neither of course, do the type labels in SAS output, for example, but at least in many instances texts and papers do contain algebraic descriptions of those hypotheses. (Obviously, though, not for Type III for Dallal's data!) If my ignorance shows through, I will happily be corrected.

Does weighting mean a generalization of Yates' method? That method calculates an SS(A) as

$$\mathrm{SSA}_w = \sum_{i=1}^{a} w_i \left[\frac{1}{b} \sum_{j=1}^{b} \bar{y}_{ij\cdot} - \sum_{i=1}^{a} w_i \left(\frac{1}{b} \sum_{j=1}^{b} \bar{y}_{ij\cdot} \right) \Big/ \sum_{i=1}^{a} w_i \right]^2 \qquad \text{(a1)}$$

for

TABLE A1: Four Models Used by BMD P4

Sums of Squares In all 4 models SS(A) = 59.115 SS(A * B) = .282	I cells, (A × B) Equal weights	II 27 cells Equal weights	III I, but "cell weights are proportional to cell sizes"	IV This "adjusts cell weights of" III so that "the estimation procedure will be equivalent" to I
SS(B)	.788	.571	.571	.788
SS[C(A * B)]	—	2.571	—	2.571
Residual	2.571	—	2.571	—

$$\frac{1}{w_i} = \frac{1}{b^2} \sum_{j=1}^{b} \frac{1}{n_{ij}} . \qquad \text{(a2)}$$

It is known for crossed classifications with all-cells-fitted data that SAS Type III is equivalent to Yates' method. Therefore, because the results in Table A1 are the same for Model I there as in Table 1 for the standard model, one concludes that Model I is Yates. And BMDP describes Model I as "equal weights". That seems strange in view of (a1) and (a2). Then, in that context what is meant by BMDP's description of Model III "weights proportional to cell size"? In view of "equal weights meaning (a1) and (a2), is it not difficult for a data-oriented statistician using computing software to intuitively know what, in this context, is meant by "weights proportional to cell size"? I think so.

2. SAS (12/14/92) Some of the generalities in this letter are not totally correct or to the point. For example, the very specific question in Dallal's letter is essentially "Why does re-labelling e_{ijk} [of equation (1) in my paper] as $\gamma_{ij:k}$ [of equation (4)] change SS(B)?" SAS writes that this "was part of a lengthy discussion about 15 years ago" and that a convenient reference for some of them" is Searle (1987). I don't believe either of these contentions.

The letter does, however, have what I see as a very important sentence; the last sentence of its fourth paragraph, namely "It is the requirement [presumably for Type III sums of squares—

S.R.S.] that the contrast(s) for B be orthogonal to *both* the contrasts for A∗B and C(A∗B), which accounts for the difference in sums of squares between the two analyses." No doubt this is correct, but it is not very informative in describing that difference. Nor does it seem to provide any help in explaining why the sums of squares in Table 1 do not seem to jibe with the question in SAS's letter: should "the sum of squares for the balanced factor be the same whether or not the within-cell variability is included in the model as an extra effect, as opposed to being left to" residual? In the penultimate sub-section of my paper (titled 'Unbalanced Data') it is pointed out that within each level of the A-factor the data are balanced with respect to C across the levels of B, and in Table 1 SS(A) is the same in both models. Contrarywise, within each level of the B-factor data are not balanced with respect to C across levels of A; and in Table 1 SS(B) is different in the two models. At least for this way of looking at balance within an unbalanced data situation, it is the balanced locale where the sums of squares are the same, and the unbalanced one where they are not.

3. SPSS (2/16/93) After acknowledging a program error and correcting it, they get it correct: SS(B) for the two models should be, and is, the same.

4. STATA (11/27/92) Their output for SS(B) with the C-model is exactly the same as that of SAS Type III, which I consider is wrong.

5. SYSTAT (12/14/92) They believe the re-labelling done by Dallal should not affect the sums of squares other than re-labelling what is residual in the standard model as SS[C(A∗B)] in the C-model. Because that model has zero degrees of freedom for residual their package will print practically nothing.

A BRIEF CONCLUSION

This is a conclusion, maybe a prayer, but not a summary! I think that, in general, software houses need to provide clearer, more detailed and especially more specific descriptions of what their calculations are. It is true that software developers are entitled to feel that they should not have to write textbooks. But it is also true that computing usage is getting easier, cheaper, faster and more widespread, with statistical novitiates making more and more use of complicated procedures. Anything we can all do to guard against ridiculous use of these procedures has got to be worthwhile.